Project Presentation

Data Science with HR Data

by Shellena Chen & Lang Wang

Project Overview

We used external Java libraries **Tablesaw** and **Smile** to conduct data visualisation and machine learning. The analysis is conducted using a human resource dataset on Kaggle.

Data Visualisation

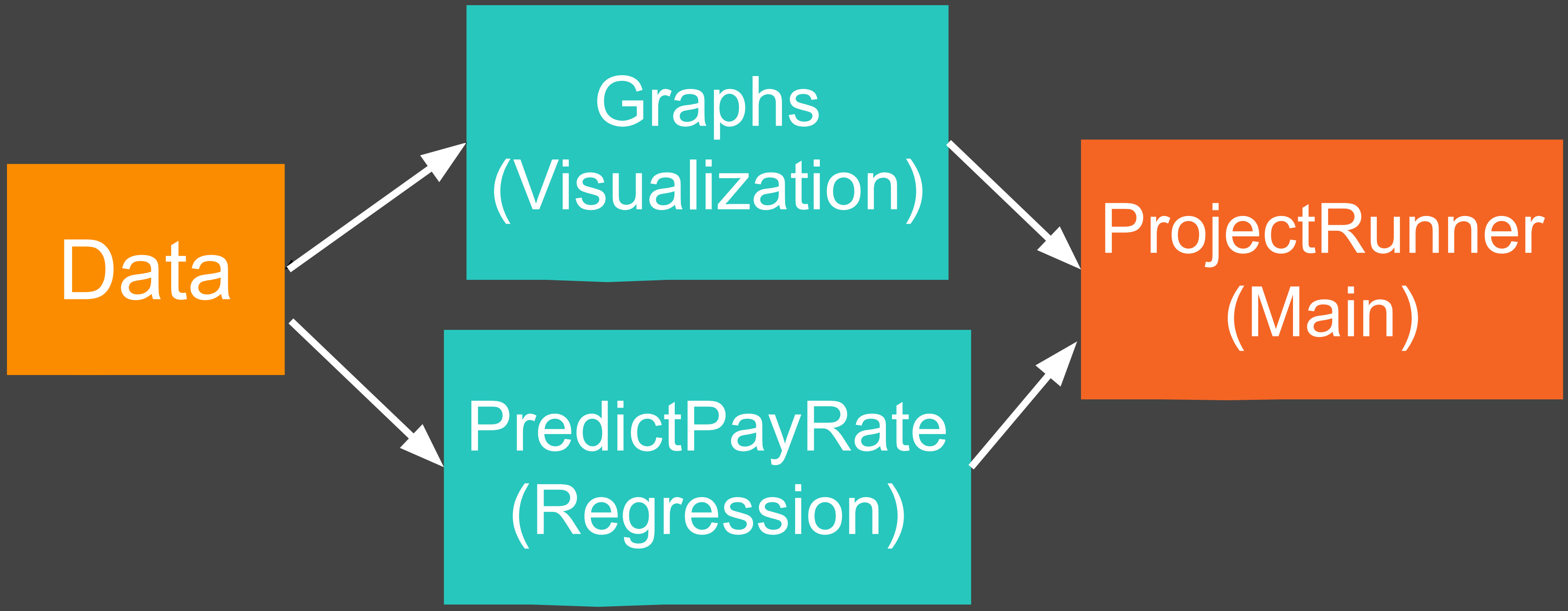
We have plotted:

- pie charts that examine the distribution of employees by gender and departments
- a histogram that examines the distribution of employees' performance; and
- box plots for employees' pay rate by citizenship status and position.

Machine Learning

We built a linear regression model using gender, employee position and employee performance score to predict an employee's pay rate. After predicting the employee's pay rate, we also evaluate the model by computing the adjusted R-Squared and R-Squared metrics.

Project Design



Data

Graphs (Visualization)

PredictPayRate (Regression)

ProjectRunner (Main)

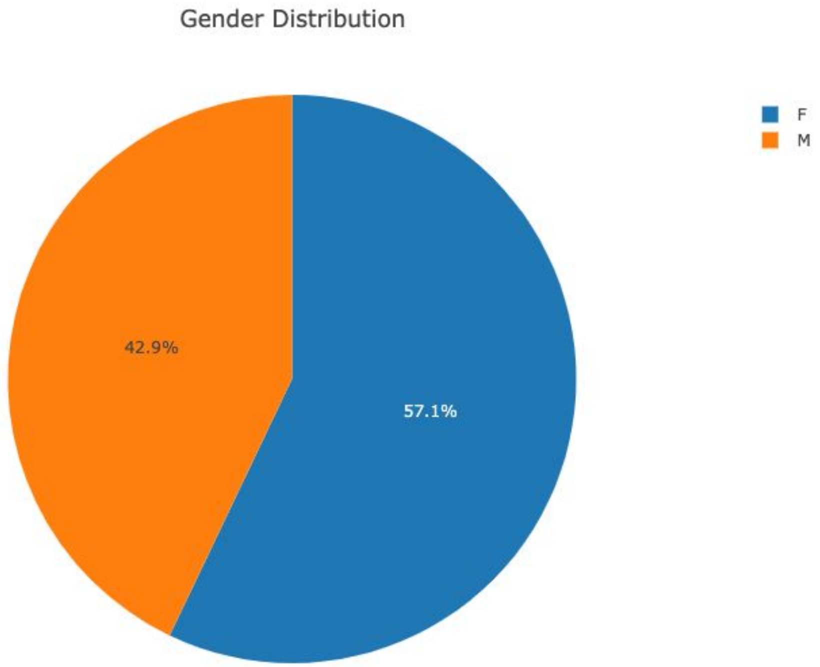
Data Exploration

We wrote various functions in the Data class to facilitate data exploration.

For instance:

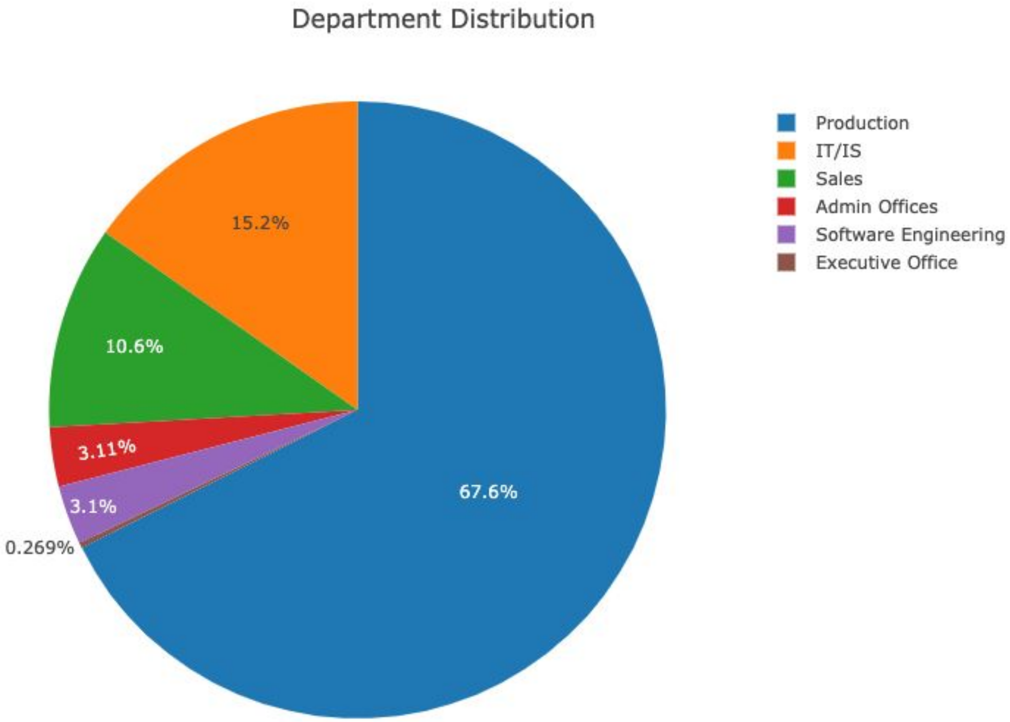
- structure of the data
- shape of the data
- column names
- first/last n rows of the data
- mean/max/min/median values of numerical columns
- correlation between columns

Data Visualization



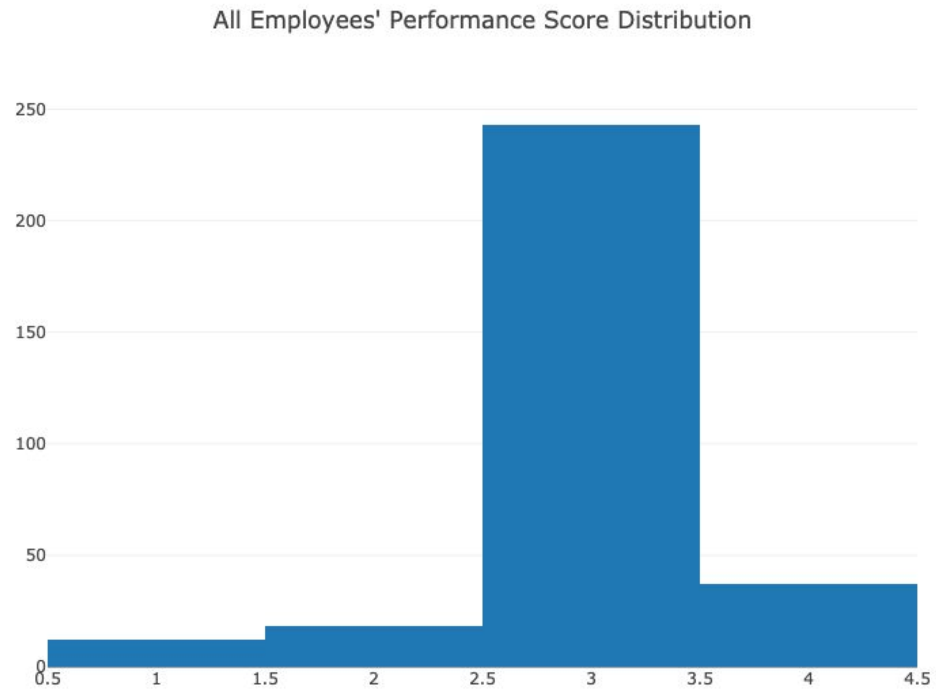
About 57% of the employees are female, and 43% male.

Data Visualization



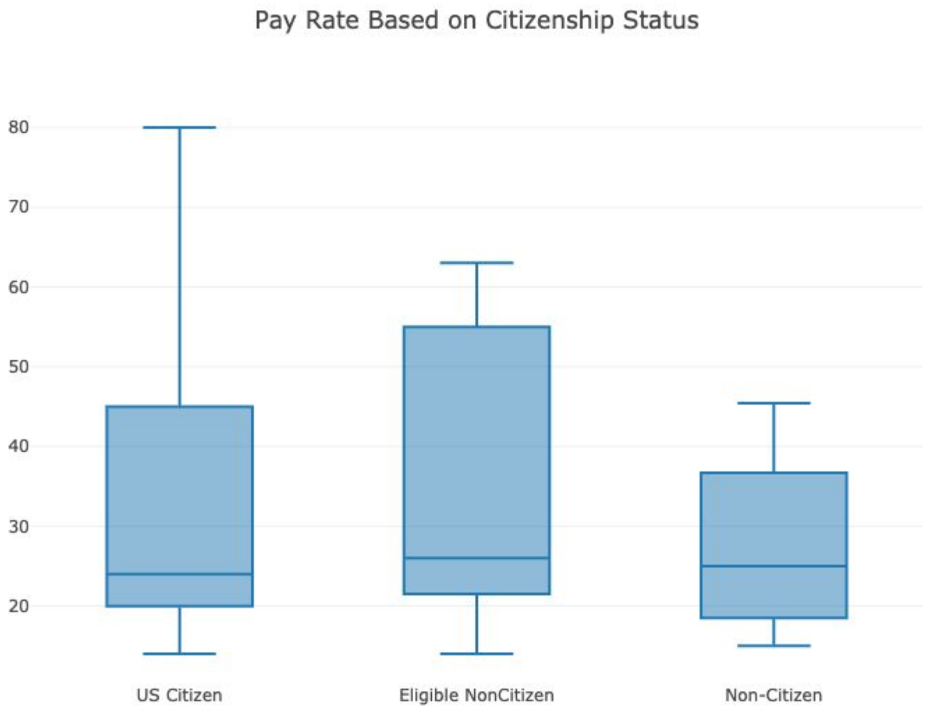
Production is the largest department.

Data Visualization



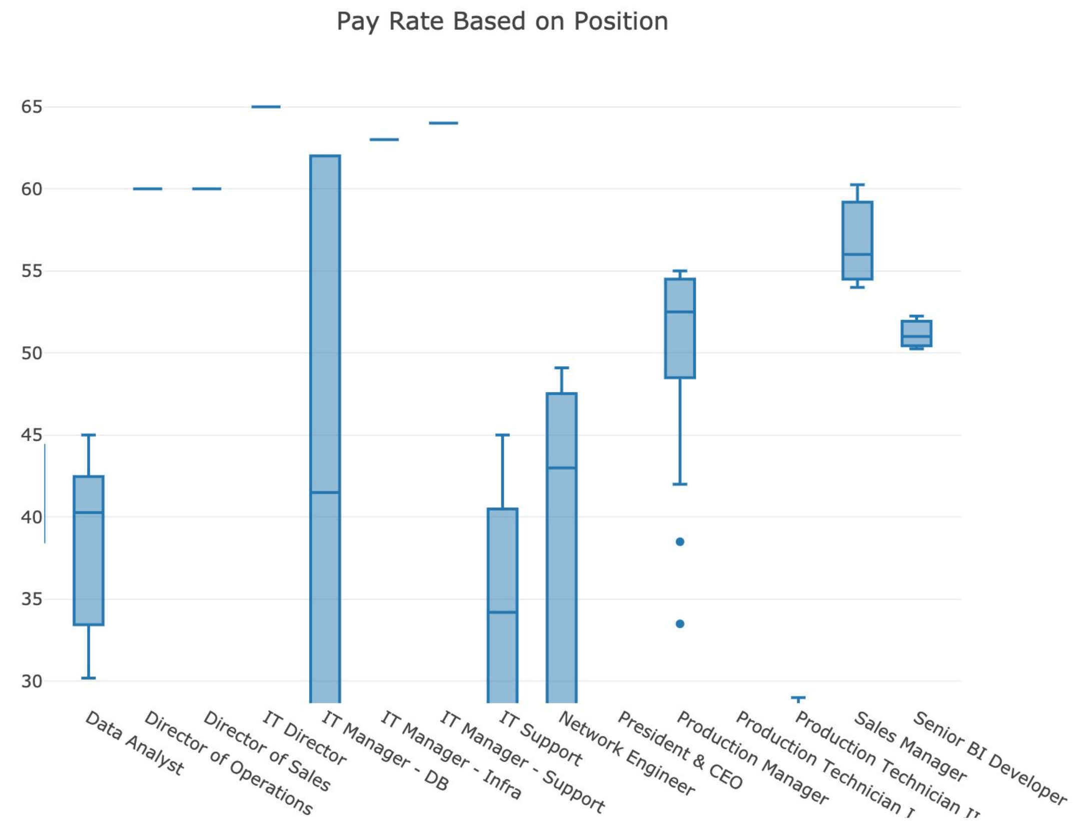
Most employees have a performance score of 3.

Data Visualization



Non-citizens have the smallest variance in pay rate, and U.S Citizens have the largest variance.

Data Visualization



Director level positions get paid more than other positions. IT Managers have the largest variance in pay rate.

Machine Learning - Linear Regression

We built a linear regression model to predict employee’s pay rate.

1. We picked gender, employee position and employee performance score as our *independent variables* and employee's pay rate as our *dependent variable.*
2. We splitted the data set into training and testing sets with a 70%-30% split, then we fitted the model with the training set and predicted pay rate with the testing set.
3. We also wrote a function to ask for user’s input on gender, employee position and performance score, then output the predicted employee pay rate by using our fitted model.
4. Lastly, we evaluated the model using the adjusted R-Squared and R-Squared metrics.